ICGC PCAWG
RNA-Seq Re-alignments

André Kahles

Postdoctoral Fellow
Gunnar Rätsch Lab
MSKCC, New York

February 23, 2015

Mission of ICGC PCAWG-3: Integration of transcriptome and genome

- Scientific mission
  - Characterize cancer transcriptome alterations that contribute to cancer phenotypes and associate the underlying genomic alterations
    - Identify somatic/germline variants that lead to expression or splicing changes
    - Identify cancer-specific expression (e.g., novel lncRNAs), splicing, transcript processing or gene fusions
    - Understand patterns of allele specific expression
    - Gain insights into regulation of RNA processes

- Data deliverables
  - Provide a unified analysis of PCAWG RNA-Seq data (alignments, expression, splicing, fusions, …)
  - Identify and quantify transcriptome-level cancer genome alterations

(as of January 2015)

PCAWG-3 Analyses/Deliverables

Analysis	Tool/Method	Group
Gene quantification	Htseq-count	Brazma, Rätsch
	RSEM	Gerstein/ENCODE pipeline
Exon quantification	bedtools	Stuart, Brazma, Rätsch
Junction quantification	Custom	Stuart, Rätsch
RNA editing	Custom tool	Hery Yang, Jun Wang
Allele-specific expression	UNCeqR	Wilkerson
Alternative promoters	Custom R/Bioconductor	Tan
lncRNA	Cufflinks, Cuffdiff	Brazma
	Cufflinks, HTSeq	Samir Amin, Chin
Transcript structure	RNA Architect	Shlien
	MiTie	Rätsch
Alternative splicing	JuncBASE	Meyerson
	DEXSeq, SwitchSeq	Brazma
	Cufflinks, Diffsplice	Guinney, Sage Bionetworks
	SplAdder/rDiff	Rätsch
Fusion transcripts	FusionSeq	Gerstein
	deFuse, Fusion Map	Brazma
	BreakTrans	Ken Chen
	FusionSeq v2	Sboner
Expression/Splicing QTL	LMM (Limix)	Rätsch

Data Overview

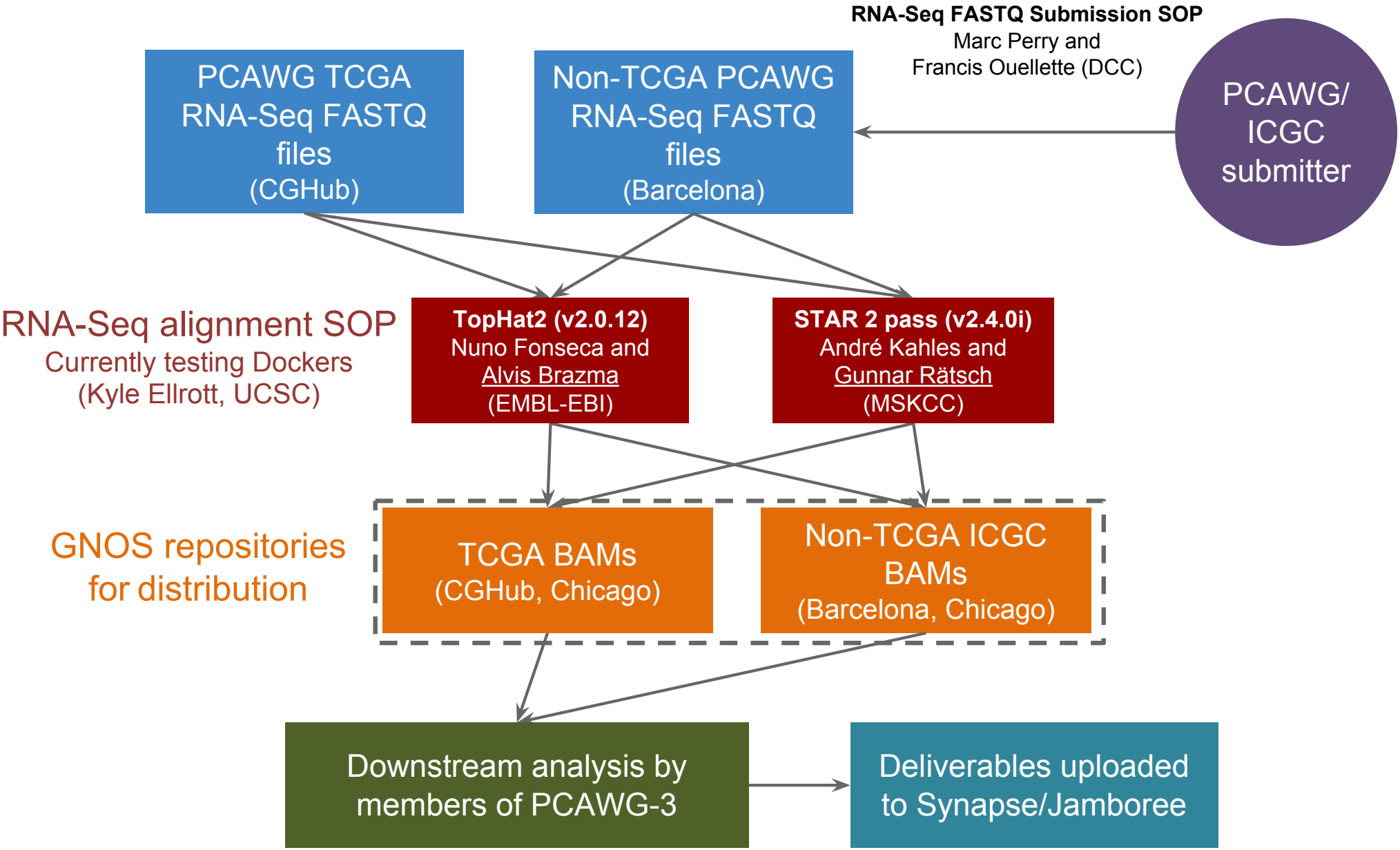
Source	Tumor type	Number of samples
ICGC	Ovarian Cancer – Serous cystadenocarcinoma	63
	Malignant Lymphoma	48
	Chronic Lymphocytic Leukemia	5
	Pancreatic Cancer – Ductal adenocarcinoma	4
TCGA	Breast Cancer	106
	Lung Adenocarcinoma	58
	Liver Hepatocellular carcinoma	56
	Uterine Corpus Endometrial Carcinoma	54
	Lung squamous cell carcinoma	51
	Head and Neck Thyroid Carcinoma	51
	Kidney Renal Clear Cell Carcinoma	46
	Colon Adenocarcinoma	45
	Skin Cutaneous melanoma	37
	Acute Myeloid Leukemia	36
	Kidney Renal Papillary Cell Carcinoma	34
	Brain Glioblastoma Multiforme	30
	Ovarian Serous Cystadenocarcinoma	23
	Prostate Adenocarcinoma	21
	Brain Lower Grade Glioma	20
	Cervical Squamous Cell Carcinoma	16
	Rectum Adenocarcinoma	15

- currently ~900 RNA-Seq samples with matched WGS
- ~500 additional samples currently collected
- 80% paired-end libraries
- read lengths 45-76nt
- library sizes 14-155M reads

Currently 743 processed and available for analysis.

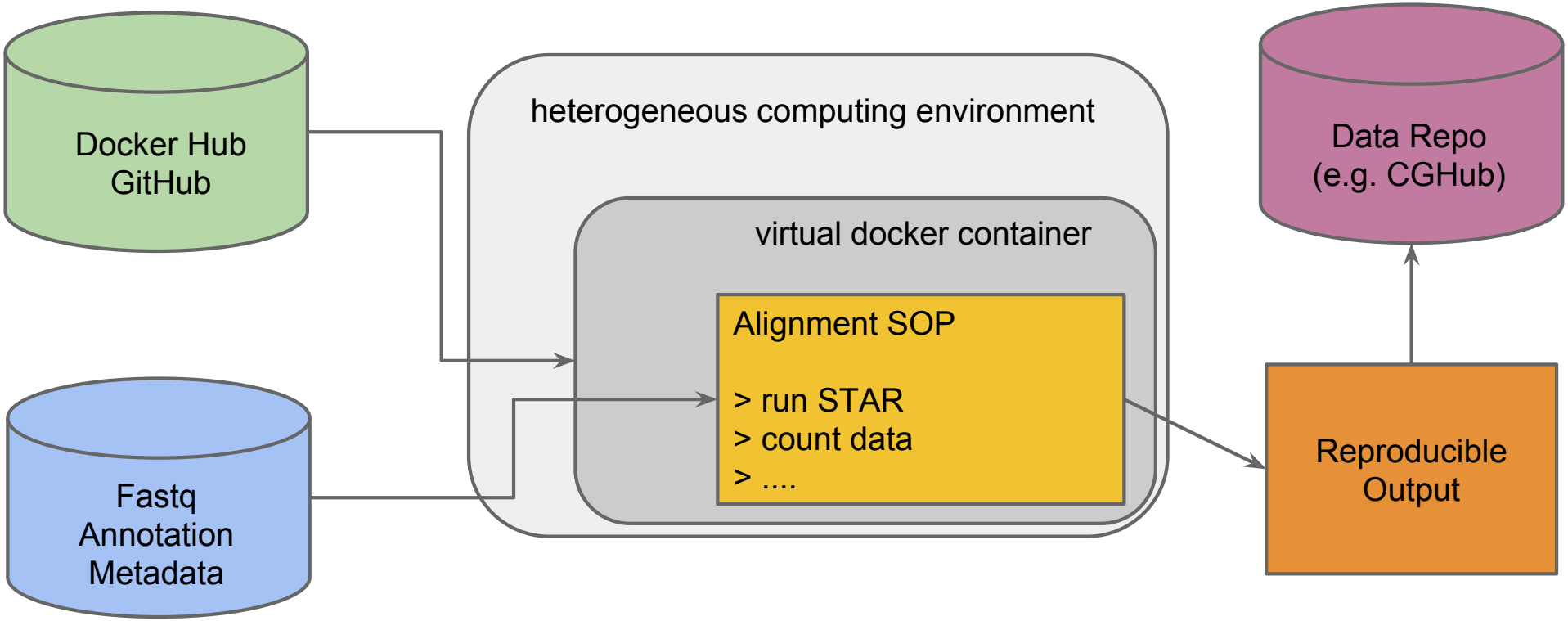
Marc Perry, Francis Oullette (DCC/OICR)

Current Alignment Setup



Courtesy of Angela Brooks (Broad Institute)

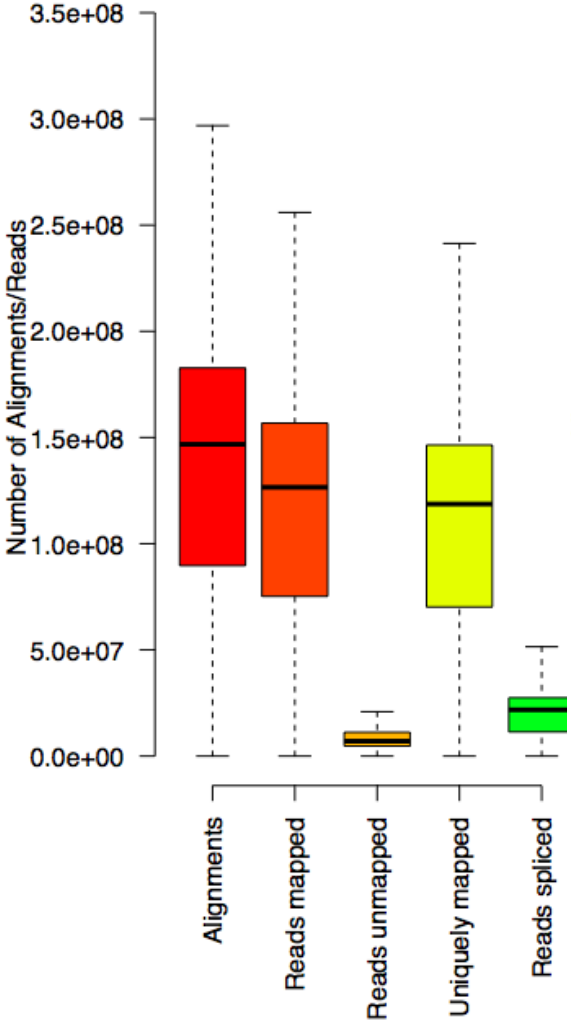
Code contained in docker images

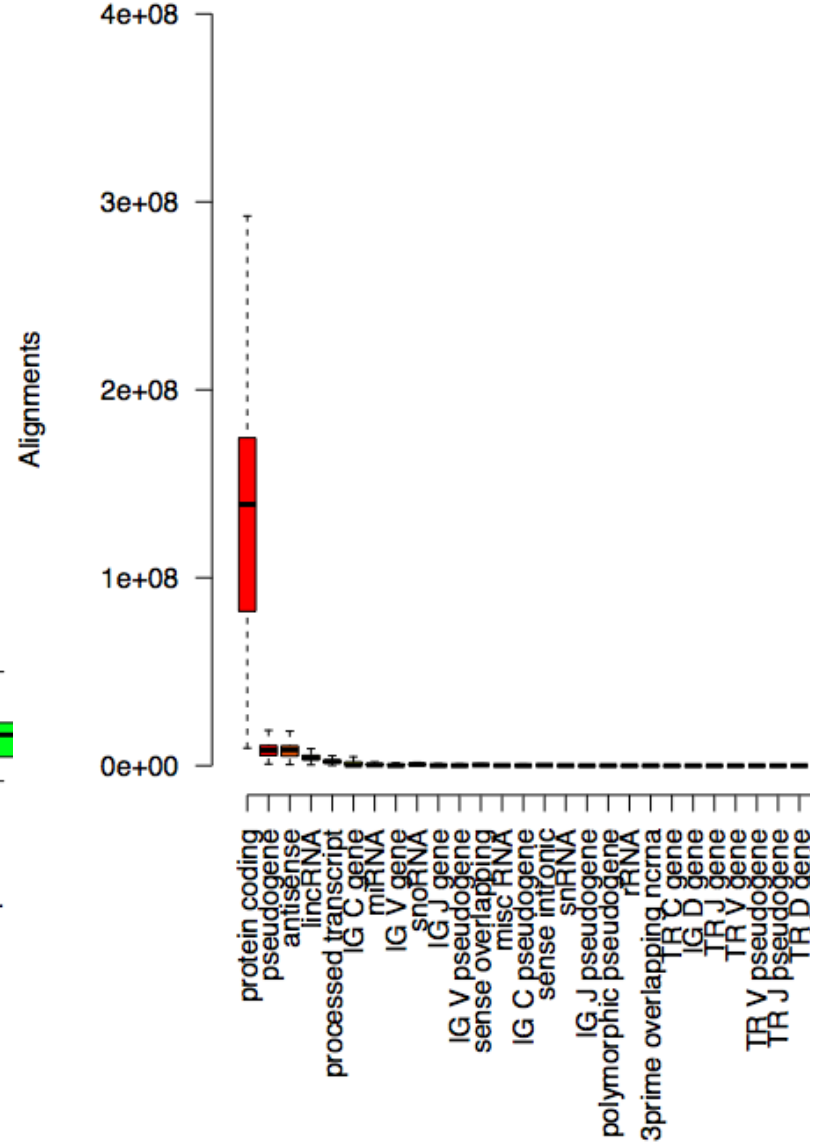


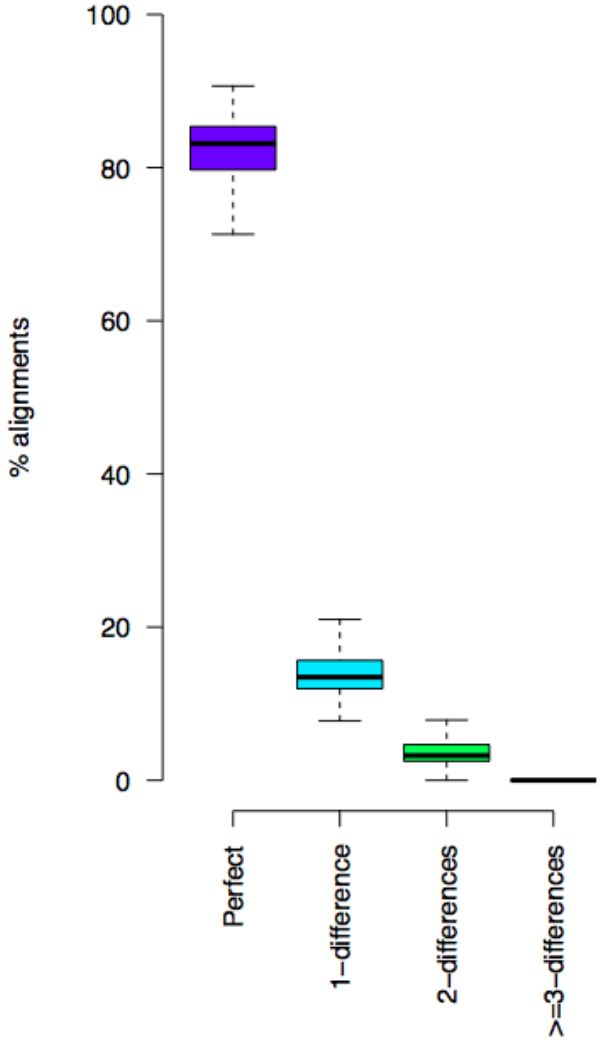
- No dependencies on a specific compute environment
- Guaranteed reproducibility
- Improved granularity and scalability / distribution

URL: https://github.com/ucscCancer/icgc_rnaseq_align

Alignment Statistics



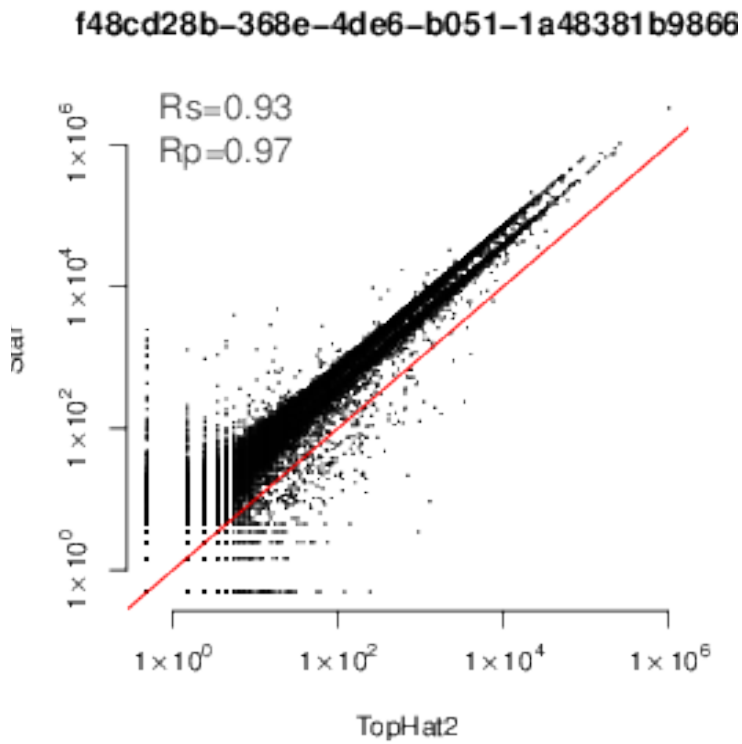


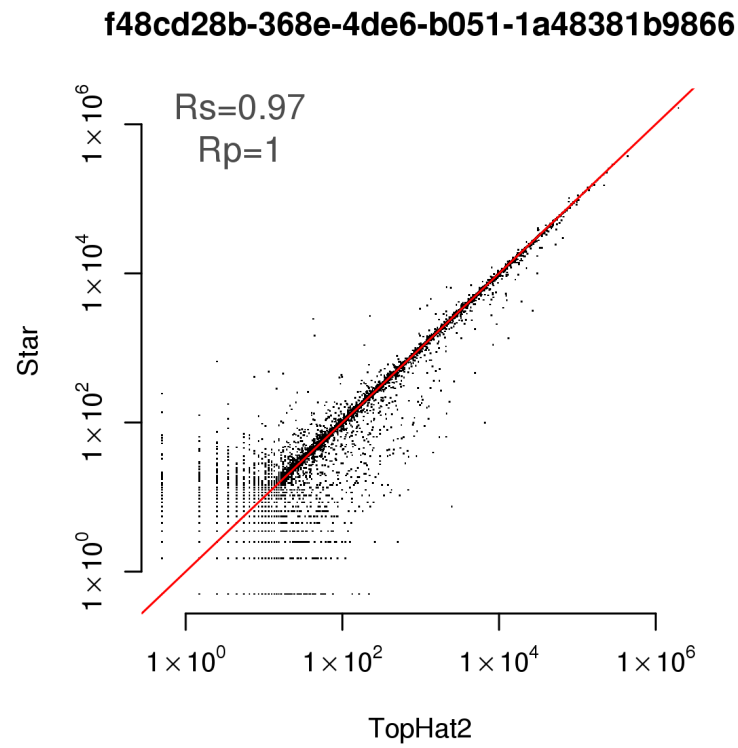


Courtesy of Nuno Fonseca (EBI)

Quantification (STAR *vs* TopHat2)

- Gene level quantification comparison between STAR and TopHat2 pipelines
- Using two pipelines allowed the **detection** and **correction** of issues in analysis

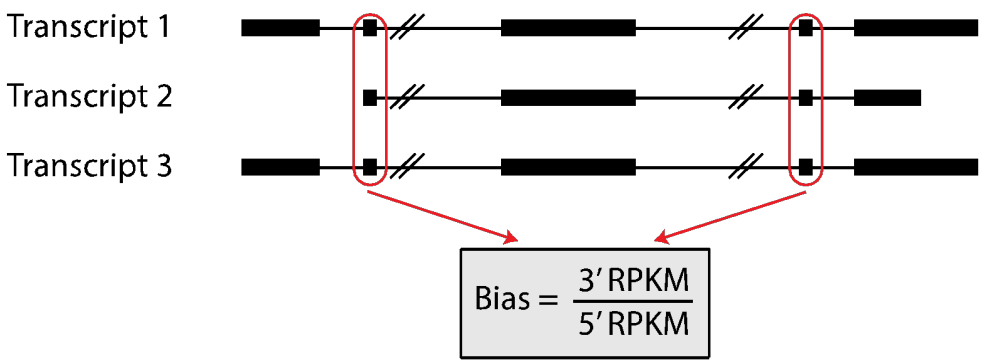




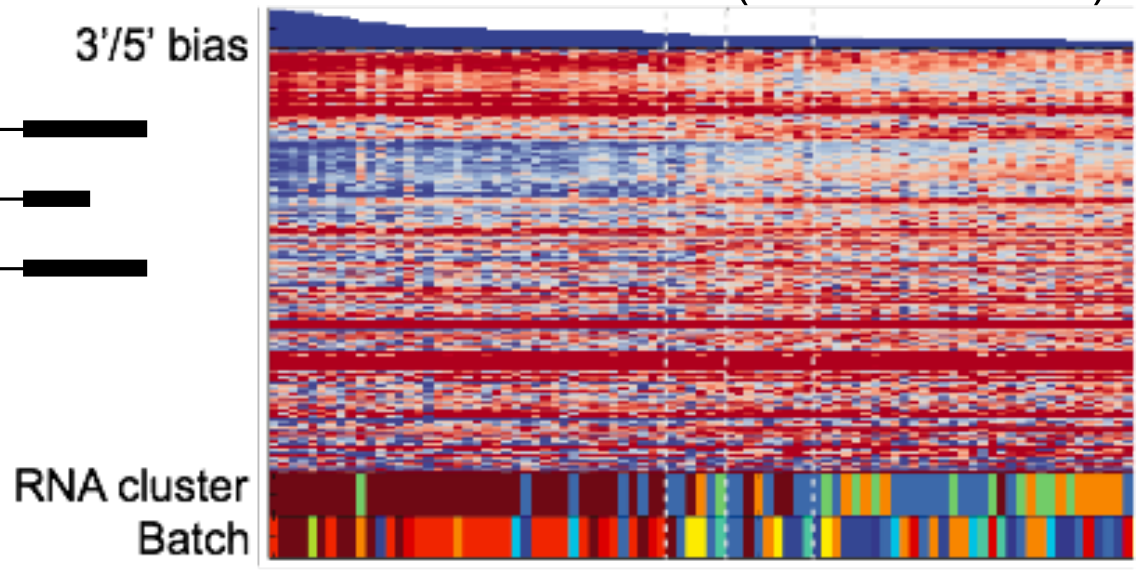
Plans to report averaged expression as “final” gene quantification.

Nuno Fonseca (EBI) & Andre Kahles (MSKCC)

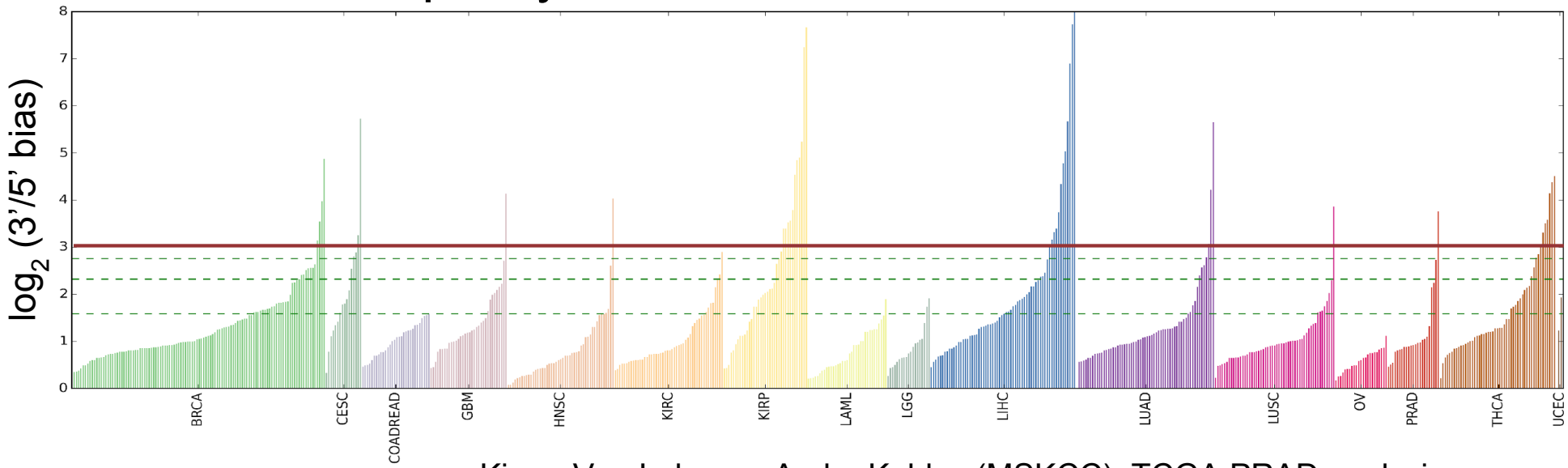
RNA QC/Degradation Analysis



PRAD (before curation)



Good overall quality of TCGA fraction of PCAWG.



Kjong-Van Lehman, Andre Kahles (MSKCC), TCGA PRAD analysis group

Further QC Efforts

Completed QC:

- FASTQ quality control (format validity, ID duplication)
- expression correlation with TCGA alignments
- concordance of expression results (STAR *vs* TopHat)
- collection of general alignment statistics

Planned/Ongoing QC:

- investigate batch effects (partially done)
- assess potential sample swaps (match germline WGS/SNPs)
- analyze read duplication

What are the ENCODE QC measures?

Unified SOP for RNA-seq

Proposal: Unified SOP for TCGA, ICGC, ENCODE, GTEx, ...

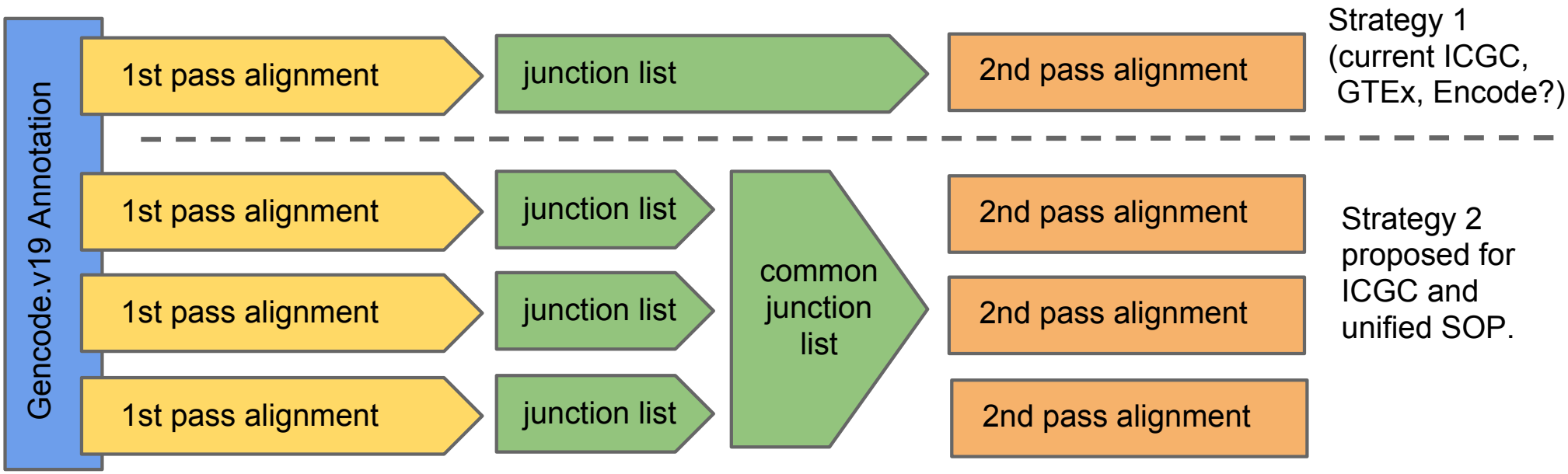
Advantages of unified SOP:

- results can be compared
- results can be integrated to be commonly used in other studies (e.g., QTL identification)
- SOP can be developed on a large-as-possible dataset

Disadvantages:

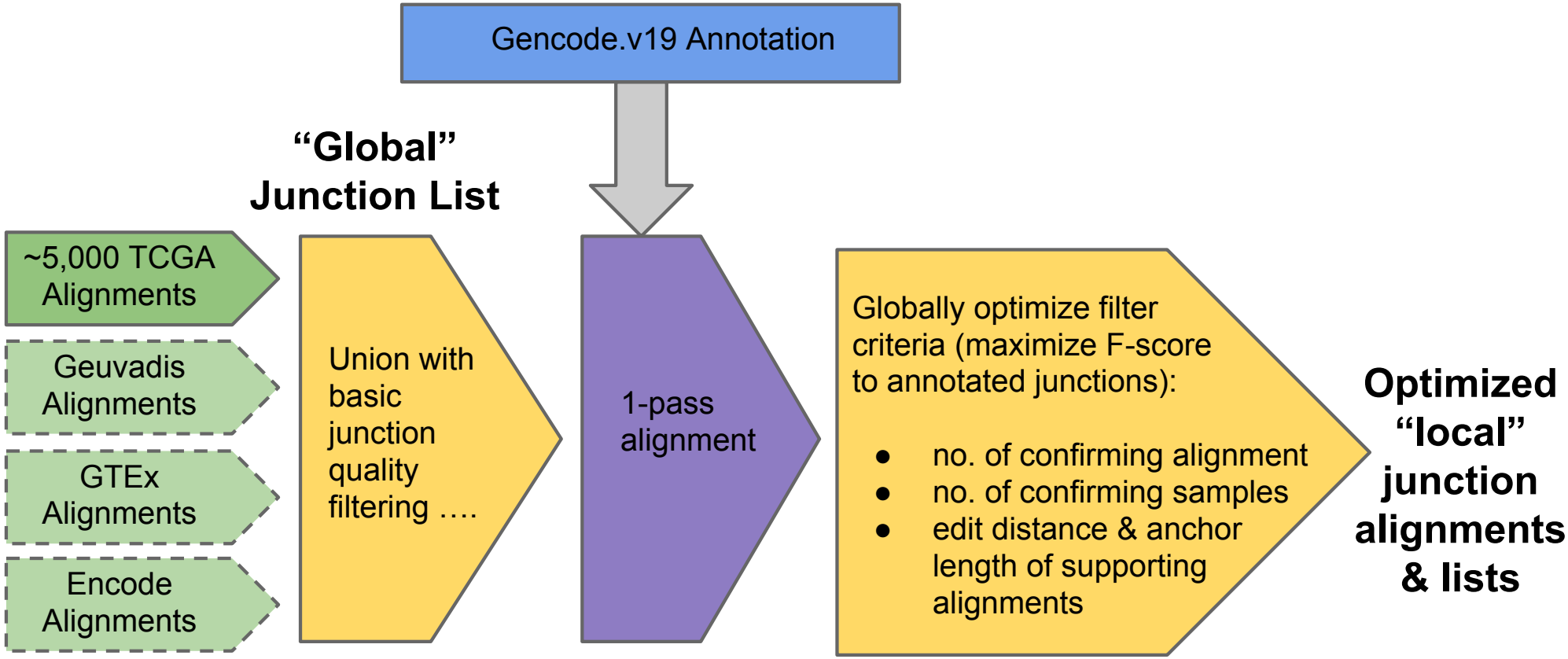
- possible effort to recreate existing results under new SOP (only 1-time effort)
- more compute for “exotic” requirements (RNA alignments usually fast)

Possible Alignment Strategies



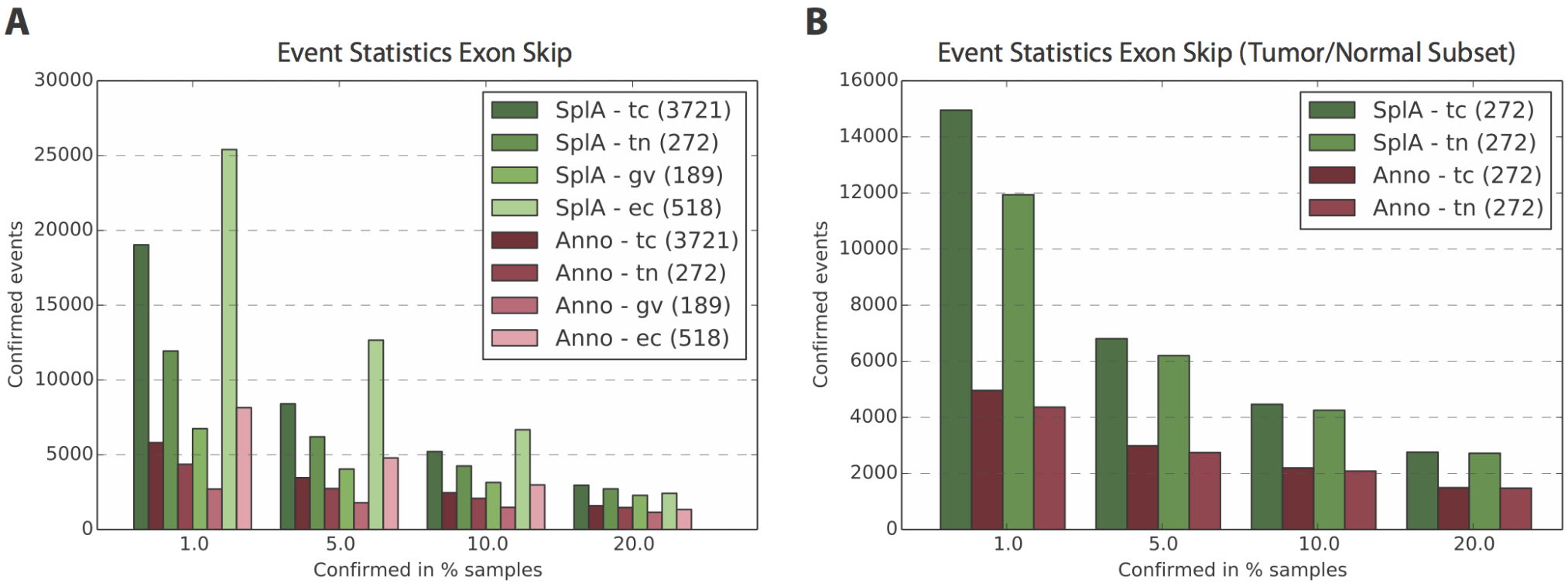
Goal: Reduce annotation bias and account for tissue-specific and tumor-specific splicing

Proposal:
2-Pass Alignments with Global Junction List



- Increased sensitivity through integrated junction list
- Improved specificity by optimized junction filtering

Tumor Specific Alternative Splicing



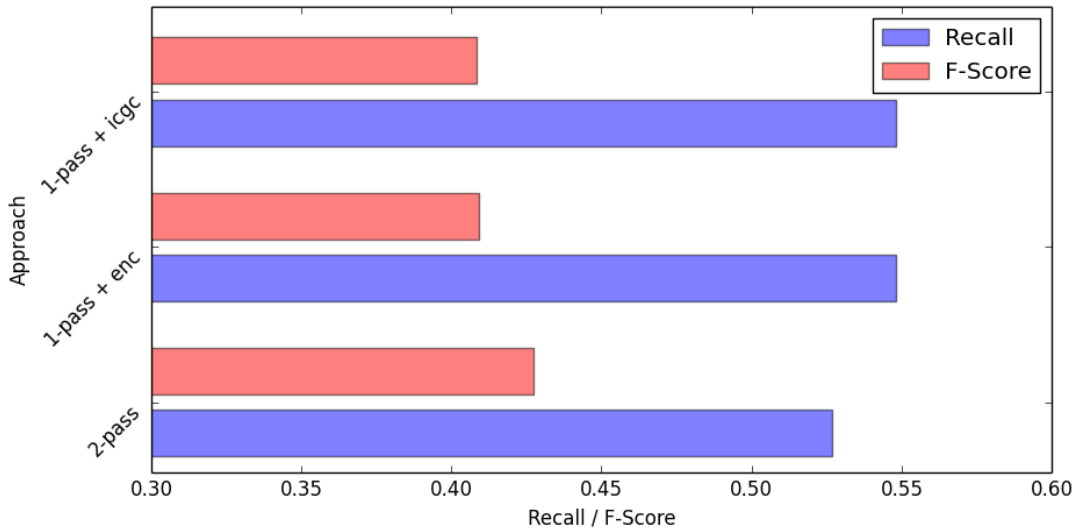
A: Exon skip events with >=5 reads per intron in TCGA/tumor (tc), TCGA/normal (tn), Geuvadis (gv) and Encode (ec). Shown are events based on the annotation and based on SplAdder (Kahles et al.).

B: Elevated degree of alternative splicing in tumor vs normal samples (sample size corrected).

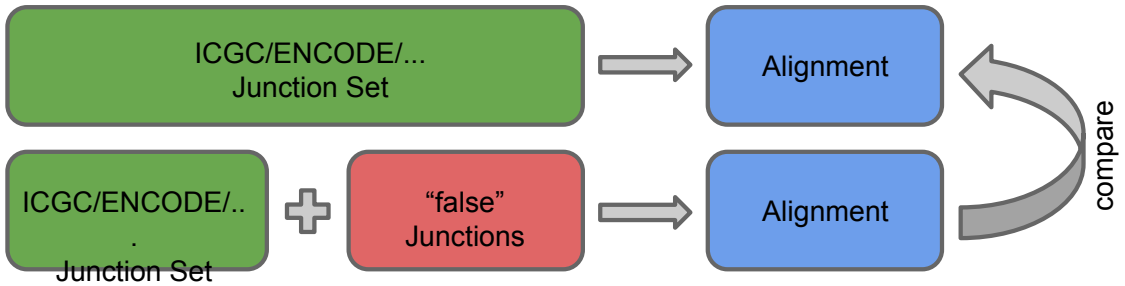
Conclusion: Both tissue- and cancer-specificity contribute significantly to splicing complexity.

Andre Kahles, PhD thesis, 2014

Junctions can improve Performance



- Using additional junctions improves recall (w.r.t. annotation), but specificity is difficult to assess



- Use "false" junctions that can not be found by 2-pass alignment to measure effect of "junction noise"

This is preliminary and needs more discussion. **Proposal to discuss and improve this idea with a small group of interested people.**

Summary Part I

- ICGC PCAWG will have ≈1,400 donors with WGS+RNA-seq
- Developed alignment SOP based on STAR and Tophat2
- Implemented in Docker for easy deployment and reproducibility
- Two workflows are more effort, but helped identifying major issues
- Basic QC analysis, what are the standards in ENCODE?
- **Proposal:** Work out a unified RNA-seq alignment SOP
  - Improved interoperability between the major projects
- **Proposal:** Alignment strategy based on a large junction list with subsequent stringent alignment filtering
  - Best combination of sensitivity & specificity

If there is time: there is a second part with analyses of 12 TCGA cancers:

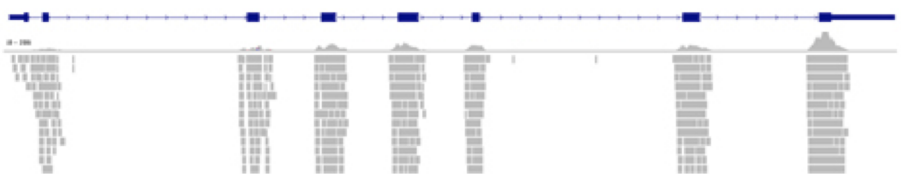
- Cancer-specific splicing
- expression and splicing QTLs based on exome & RNA-seq data

Part II: TCGA Analysis Across 12 Cancers

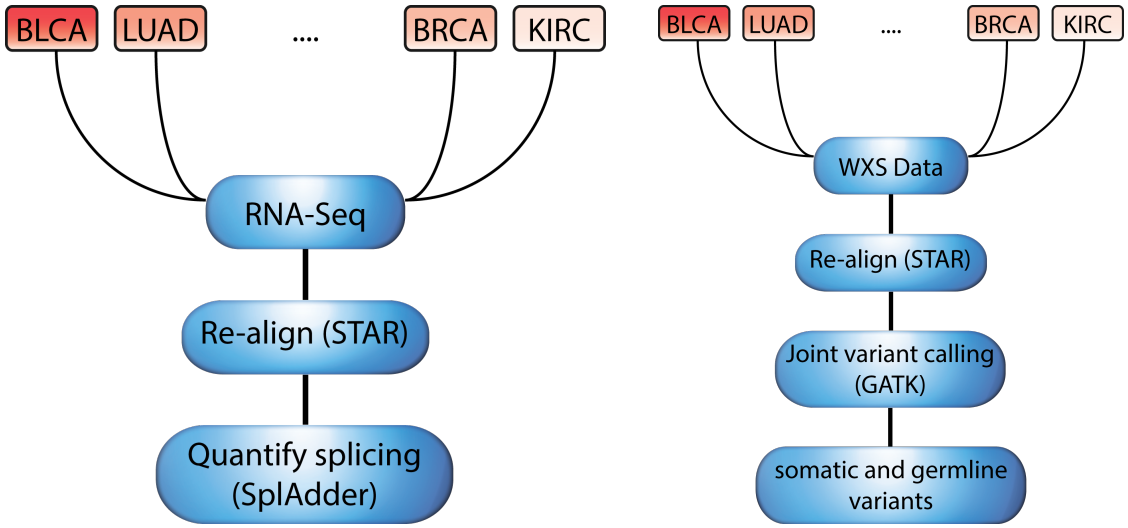
1. Identify cancer-specific splicing patterns
2. Identify variants regulating splicing in the same gene (cis)
3. Identify variants regulating splicing in other cancer genes (trans)

TCGA provides RNA-seq and matching exome data

- RNA-seq ⇝ Find & quantify splicing events
- Exome ⇝ Identify variants in exons & flanking intronic regions



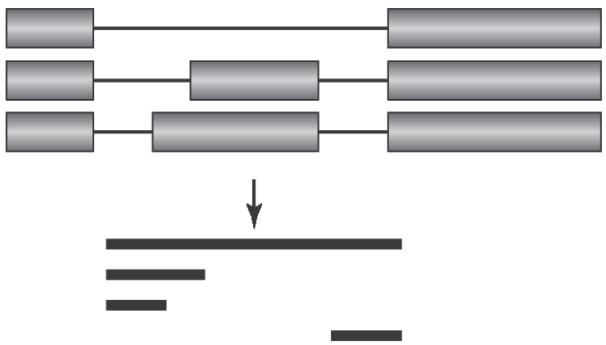
Problem: Non-uniform processing (alignments & variant calling)



Kjong-Van Lehmann, Andre Kahles, Gunnar Rätsch, *et al*. (MSKCC)

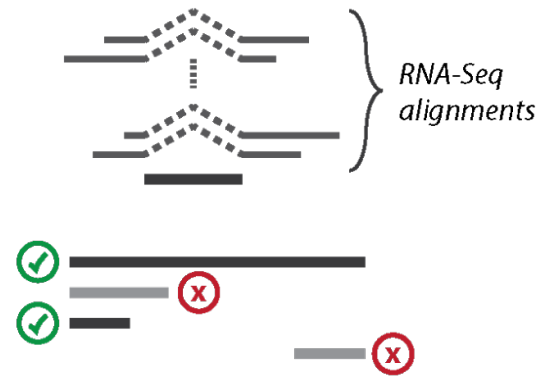
Detection of Cancer Specific Splicing

A



Compile list of all annotated and newly observed introns

B



Mark introns as expressed in a sample if supported by >= 10 RNA-Seq reads

C

Tumor	Normal	ENCODE	Geuvadis	
✓✓✓xxx	xxxxxx	✓✓xxxx	✓✓✓✓xx	intron 1
✓✓✓✓xx	✓✓✓xxx	✓✓✓xxx	✓✓xxxx	intron 2
⋮				
✓✓✓✓xx	✓xxxxx	✓✓xxxx	✓✓✓✓xx	intron n-1
✓xxxxx	✓✓✓✓✓x	✓✓✓xxx	✓✓xxxx	intron n

Assess expression of each intron in tumor, normal and outgroup samples

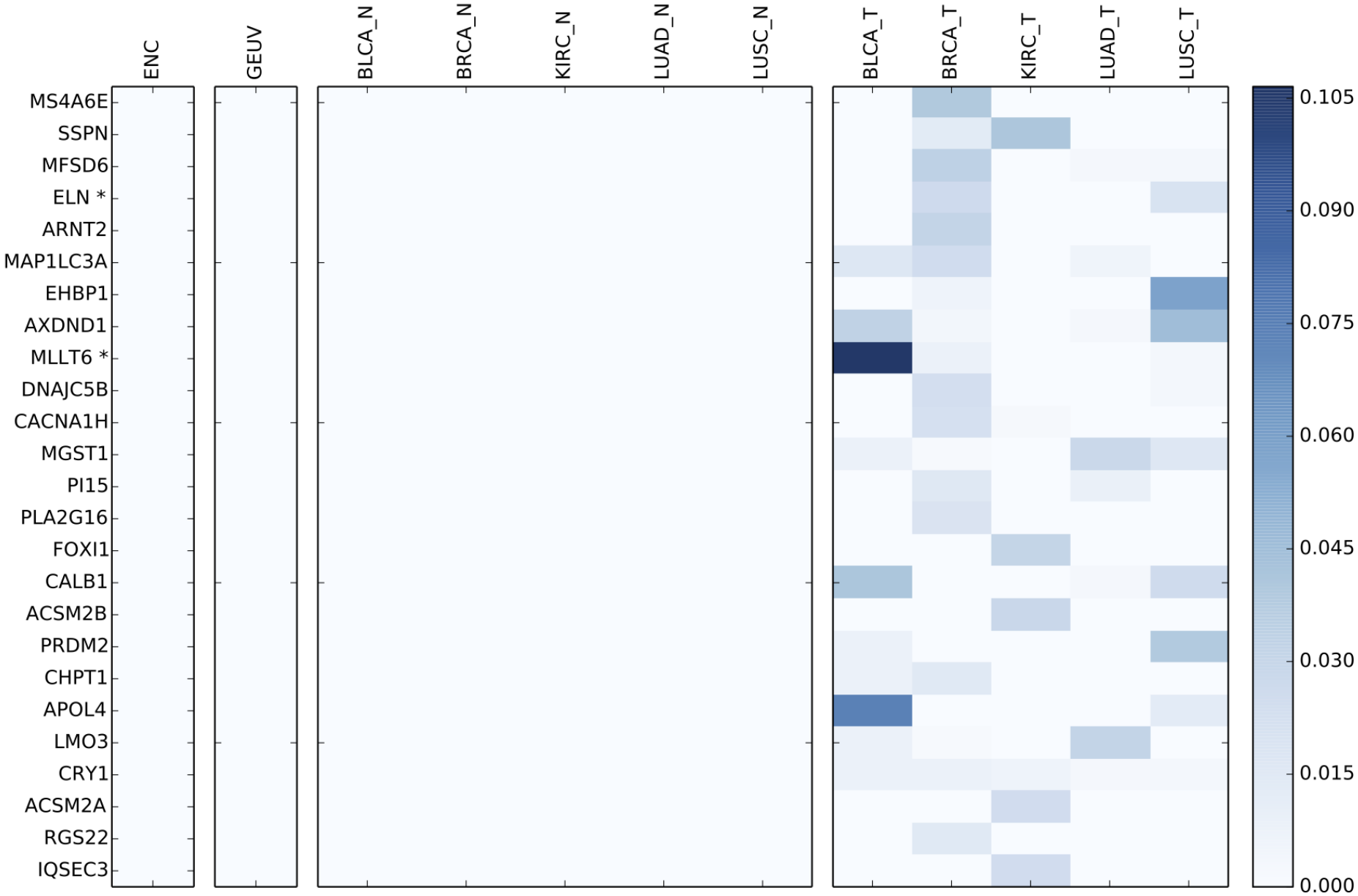
D

Tumor	Normal	ENCODE	Geuvadis	
✓✓✓✓✓✓	xxxxxx	xxxxxx	xxxxxx	intron 3817
✓✓✓✓✓✓	xxxxxx	xxxxxx	xxxxxx	intron 79
✓✓✓✓✓x	xxxxxx	xxxxxx	xxxxxx	intron 1422
✓✓✓✓✓x	xxxxxx	xxxxxx	✓xxxxx	intron 732
⋮				

Re-rank introns by the ratio of tumor and non-tumor intron evidence.

Andre Kahles (MSKCC)

Result on Tumor Specific Splicing

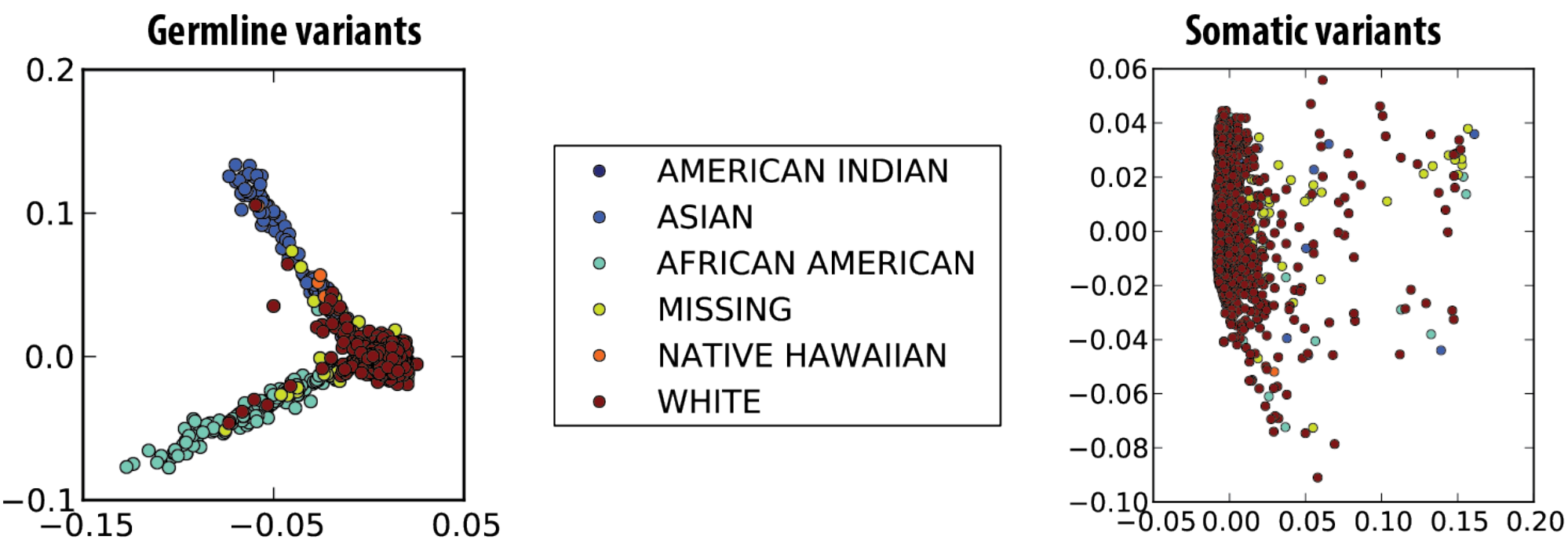


preliminary, Andre Kahles (MSKCC)

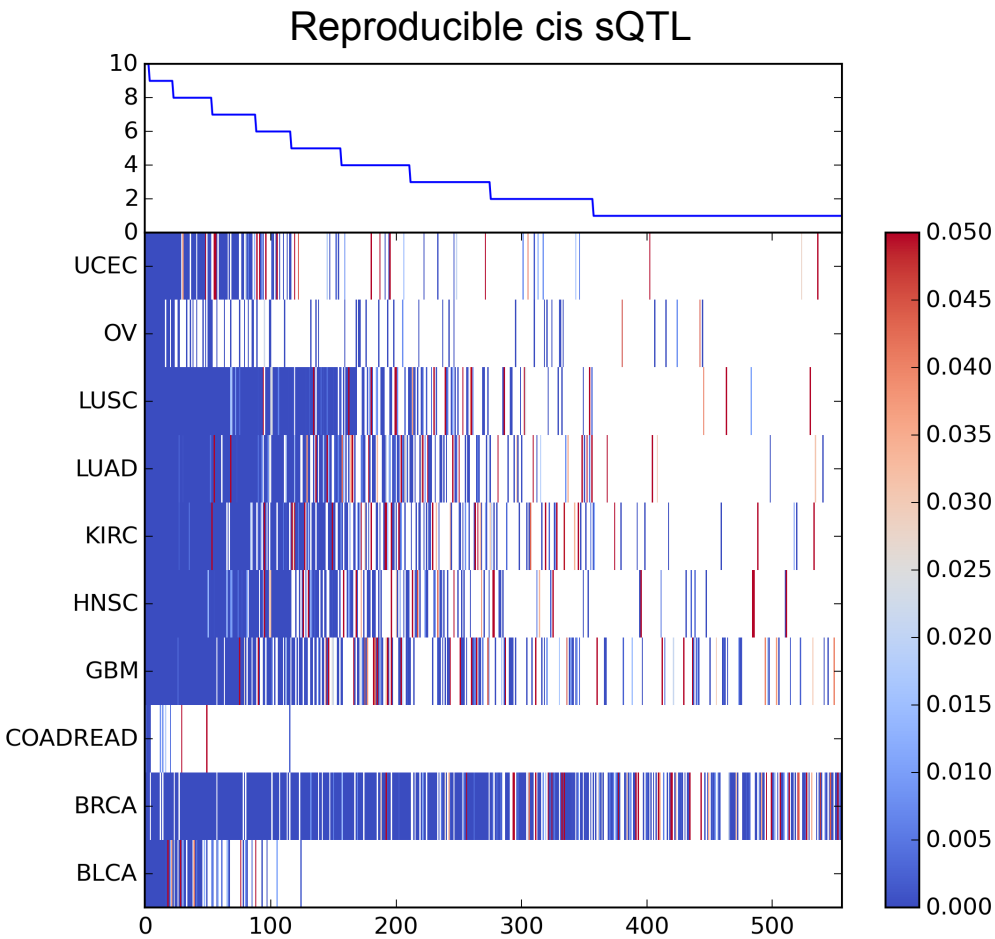
Common Variant Association

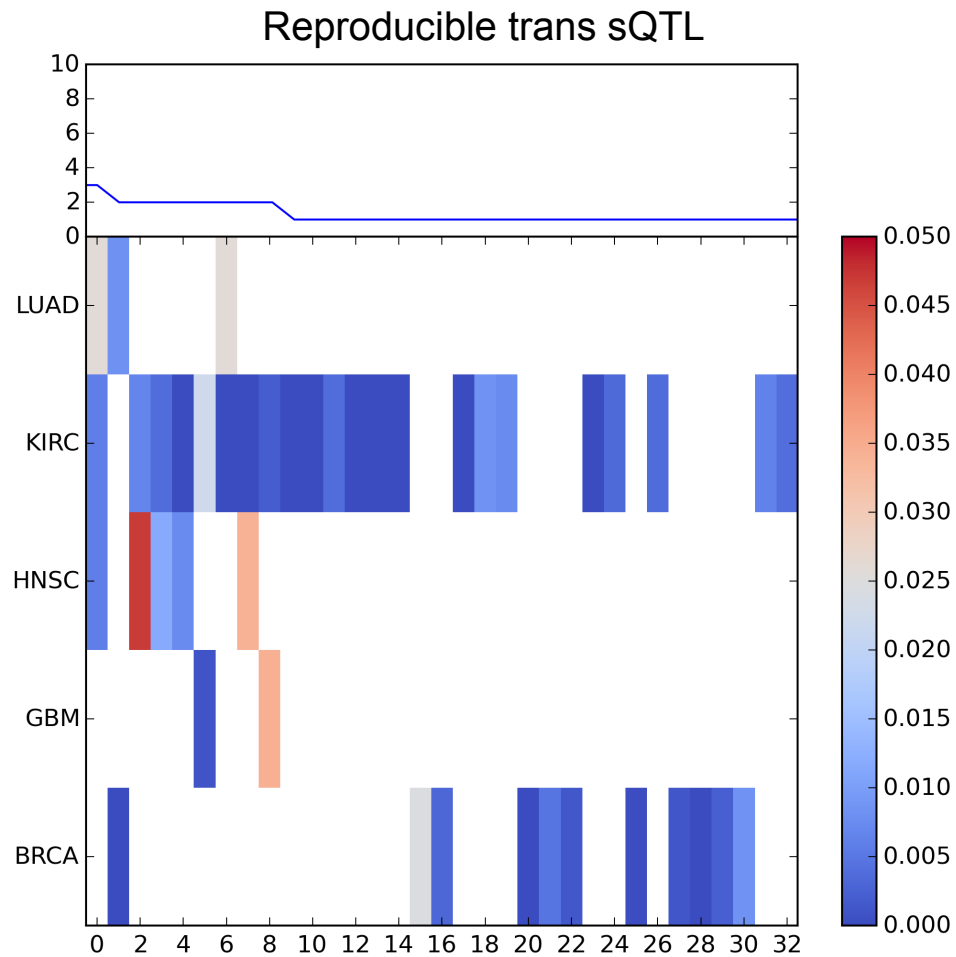
Modeling Cancer and Population Structure

- $Y = X\beta + PopulationStructure + CancerStructure + \epsilon$
- $PopulationStructure \sim N(0, \sigma_p^2 P)$ with $P = X_{\text{germ}} X_{\text{germ}}^T$
- $CancerStructure \sim N(0, \sigma_c^2 C)$ with $C = X_{\text{soma}} X_{\text{soma}}^T$



sQTL Reproducible over Cancer Types



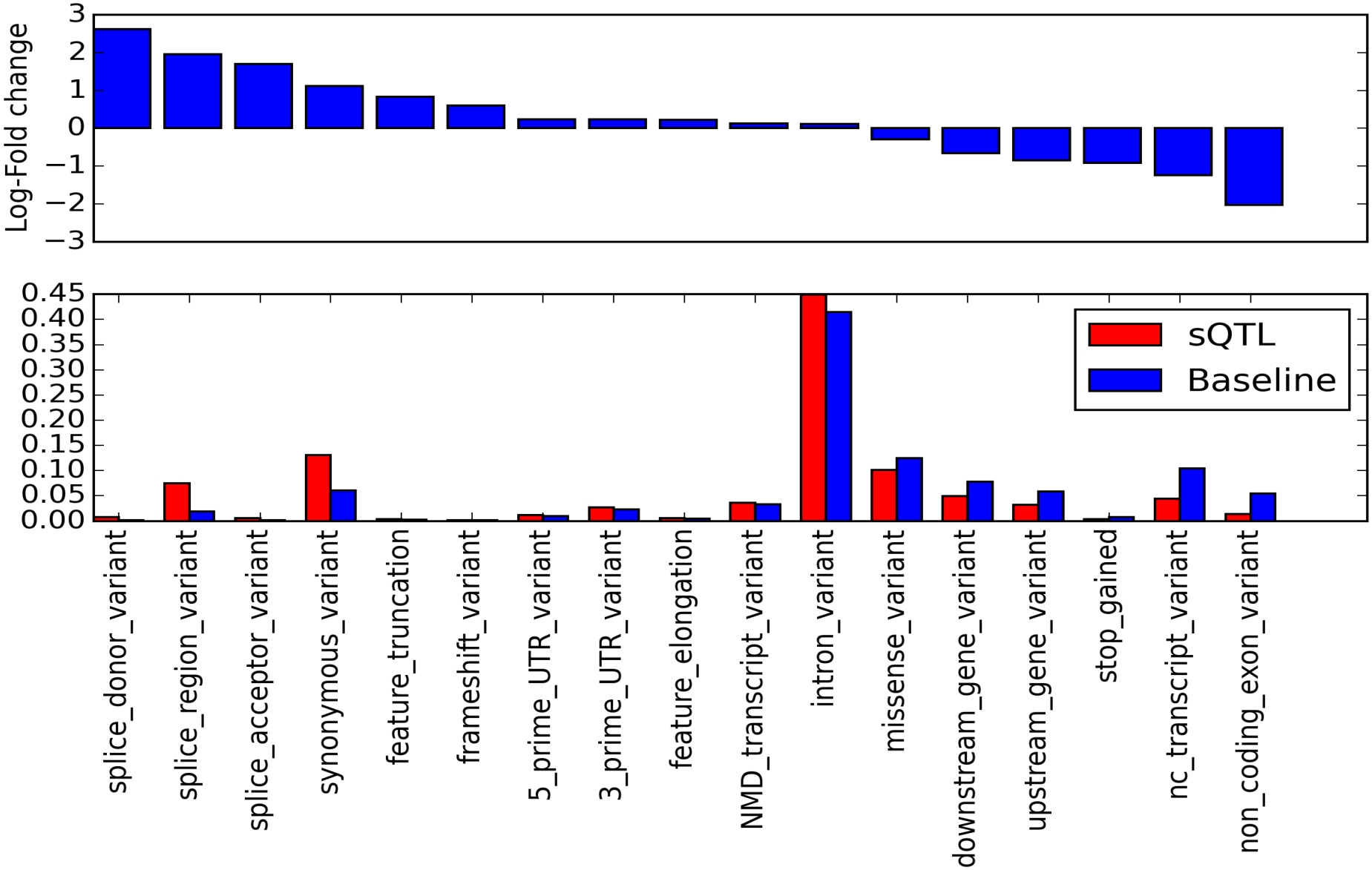


Left: Reproducibility of *cis* sQTLs: a few hundred reproduce in at least two cancer types. Number of significant sQTLs corresponds to sample sizes.

Right: Lower reproducibility for trans sQTLs, involving some of the known genes including SF3B1 ...

Preliminary, Kjong-Van Lehmann (MSKCC)

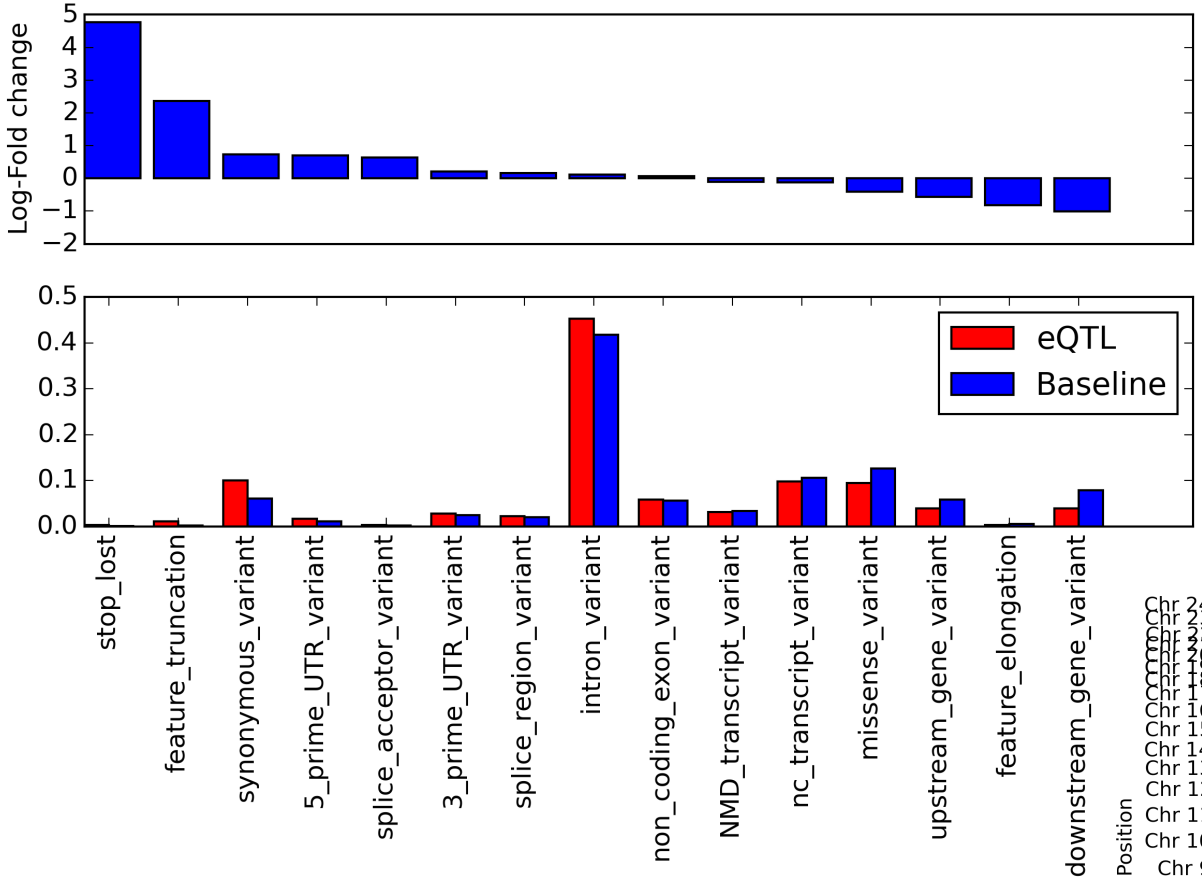
sQTL - Functional Enrichments (*cis*)



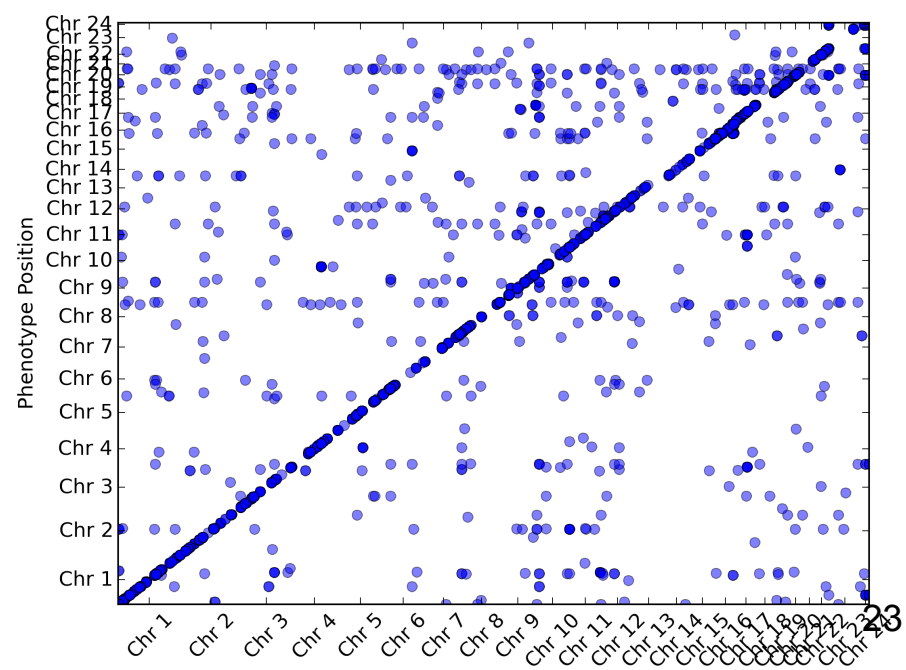
Preliminary, Kjong-Van Lehmann (MSKCC)

eQTL Results

Functional enrichment (*cis*)



Significant *trans* eQTLs



Preliminary, Kjong-Van Lehmann (MSKCC)

Part I

PCAWG-3 Group

Broad Institute
Matthew Meyerson*
Angela Brooks*
Gaddy Getz
Chandra Pedamallu
Sam Freeman
David DeLuca
Ayellet Segre
Tim Sullivan
Lihua Zou

EMBL-EBI
Alvis Brazma*
Nuno Fonseca
Lilian Greger
Mar Gonzàlez-Porta
Oliver Stegle

MSKCC
Gunnar Rätsch*
Andre Kahles
Kjong Lehmann
Natalie Davidson
Stefan Stark

Part II

University of Chicago
Zhenyu Zhang
Allison Heath
Bob Grossman

UC Santa Cruz
Kyle Ellrott
Christopher Wilks

University of Tokyo
Yuichi Shiraishi

Ontario Institute for Cancer Research
Francis Ouellette
Marc Perry

Baylor College of Medicine
Chad Creighton
Yuan Yuan

UNC Chapel Hill
Matthew Wilkerson
Katherine Hoadley

Hospital for Sick Children
Adam Shlien

MD Anderson
John Zhang
Ken Chen
Leng Han
Sahil Seth
Wanding Zhou
Xian Fan
Zechen Chong
Samir Amin
Han Liang

Peking University
Zemin Zhang

EMBL Heidelberg
Alejandro Reyes
Jan Korbel

Yale
Mark Gerstein
Anurag Sethi

Washington University in St. Louis
Reyka Jayasinghe
Venkata Yellapantula

Dana-Farber
Virginia Savova

Genome Institute of Singapore
Jonathan Goke
Tannistha Nandi
Patrick Tan

Sage Bionetworks
Yin Hu

NorthShore University Health System
Yuan Ji

BGI
Yong Hou

Weill Cornell Medical College
Ekta Khurana
Andrea Sboner

* working group leaders